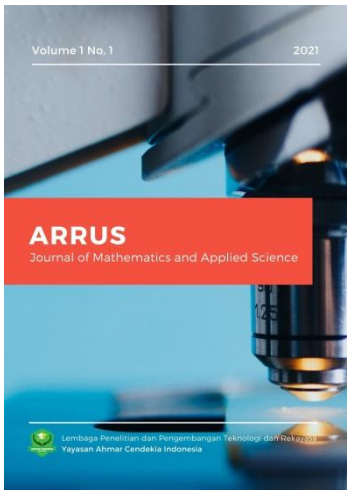
https://doi.org/10.35877/mathscience4878

OPEN ACCESS

ISSN : 2776-7922 (Print) / 2807-3037 (Online)



RESEARCH ARTICLE

SAR-to-Optical Image Translation Based on CycleGAN with Training Stabilization: A Sumatra Flood Case Study

Ahmad Imdad, Kartika Fithriasari*, & Setiawan

Department of Statistics, Institut Teknologi Sepuluh Nopember (ITS), Surabaya, Indonesia.

*Corresponding author: Kartika Fithriasari, Department of Statistics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia.

E-mail: kartika_f@its.ac.id

Abstract: This study employs CycleGAN to translate SAR into optical-like imagery in an unpaired-data scenario within flood-affected areas of Sumatra. Seven training configurations were evaluated, including the default setting, an asymmetric learning rate scheme, a combination of spectral normalization and geometric augmentation, isolated ablations of these components, and variations in the cycle-consistency coefficient. The dataset consisted of 482 Sentinel-1 SAR patches and 446 Sentinel-2 optical patches for training, alongside 276 SAR images for testing, all acquired via Google Earth Engine during the November 2025 flood. The evaluation utilized four metrics: the Structural Similarity Index (SSIM), Learned Perceptual Image Patch Similarity (LPIPS), Fréchet Inception Distance (FID), and the convergence epoch as the primary model-selection criterion. SSIM and LPIPS were calculated for cyclic reconstruction to address the lack of paired optical references. The configuration integrating spectral normalization, a reduced discriminator learning rate, and geometric augmentation achieved the fastest convergence (epoch 43) and the highest performance across all metrics (SSIM = 0.939, LPIPS = 0.025, and FID = 143.00). The ablation study revealed that the effectiveness of spectral normalization depends on the accompanying discriminator dynamics, underscoring the necessity of jointly reporting reconstruction-based and distribution-based metrics.

Keywords: CycleGAN, Remote Sensing, SAR, Unpaired Data

1. Introduction

Optical and Synthetic Aperture Radar (SAR) imagery play complementary roles in geospatial applications. Optical imagery provides easily interpretable reflectance information, whereas SAR relies on active microwave signals that penetrate cloud cover and operate independently of solar illumination (Moreira et al., 2013; Torres et al., 2012). In tropical regions, such as Indonesia, where cloud cover is consistently high, exclusive reliance on optical imagery becomes an operational bottleneck during disaster events. However, SAR imagery is difficult to interpret directly because of its grayscale representation, multiplicative speckle noise, and side-looking geometric distortion (Lee & Pottier, 2017). Consequently, acquiring a mapping from the SAR domain to the optical domain serves as a visual supplement when optical imagery is unavailable. (Wang et al., 2019).

This cross-domain mapping can be formulated within the image-to-image translation framework based on a Generative Adversarial Network (GAN) (Goodfellow et al., 2014). Whereas Pix2Pix (Isola et al., 2017) requires paired data, CycleGAN (Zhu et al., 2020) operates without paired correspondence by using a cycle-consistency mechanism. Because

https://doi.org/10.35877/mathscience4878

ISSN : 2776-7922 (Print) / 2807-3037 (Online)

strictly synchronized SAR–optical pairs are difficult to obtain for rapidly evolving disaster events, CycleGAN is a more flexible methodological choice. Numerous studies have adapted CycleGAN to SAR-to-optical settings (Reyes et al., 2019; Wang et al., 2019; H. Zhang et al., 2025); however, to the best of the authors' knowledge, few have addressed tropical flood contexts with mixed land cover such as that of Sumatra.

This study was motivated by the large-scale flash flood that struck three provinces of Sumatra in late November 2025, driven by Tropical Cyclone Senyar, with daily rainfall exceeding 400 mm. During this event, thick cloud cover rendered Sentinel-2 optical imagery unusable, whereas the available Sentinel-1 SAR imagery remained difficult for non-technical stakeholders to interpret. In such a setting, a SAR-to-optical translation that faithfully represents flood-affected areas through their characteristic visual signature would serve as an interpretable visual aid, allowing non-technical stakeholders to identify the spatial extent of the impacted land directly from translated imagery. However, a central obstacle to CycleGAN is the instability of adversarial training: an overly strong discriminator yields uninformative gradients for the generator (Arjovsky et al., 2017). Mitigation strategies include distinct generator–discriminator learning rates and Lipschitz constant control through spectral normalization (Heusel et al., 2017).

This study makes three substantial contributions to the existing literature. Firstly, CycleGAN was employed on a local unpaired dataset from flood-affected regions in Sumatra. Secondly, a systematic ablation was conducted across seven training configurations to identify effective stabilization strategies for SAR-to-optical cases. Thirdly, the results were evaluated using both reconstruction-based and distribution-based metrics, which are suitable for the absence of spatial correspondence in the test data.

2. Literature Review

2.1. SAR and Optical Remote Sensing Imagery

Optical remote sensing imagery, such as that acquired by the Sentinel-2 satellite, captures surface reflectance within the visible and near-infrared spectra and is represented as a three-channel RGB composite. This imagery is relatively straightforward to interpret visually but is severely limited by cloud cover, particularly in tropical regions, where persistent cloud obstruction can render the imagery unusable for weeks (Torres et al., 2012).

Synthetic Aperture Radar (SAR) imagery, such as that obtained by Sentinel-1, operates in the microwave spectrum using the active transmission and reception of radar pulses. SAR is independent of solar illumination and can penetrate cloud cover, making it available under conditions where optical imagery is unavailable (Moreira et al., 2013). However, SAR imagery presents three characteristics that limit its visual interpretability: (i) it is typically represented as single- or dual-polarized backscatter intensity in grayscale, lacking color information; (ii) the coherent nature of radar waves introduces multiplicative speckle noise; and (iii) the side-looking acquisition geometry leads to distortions like foreshortening, layover, and radar shadow, which do not occur in optical imagery.

These complementary strengths and limitations motivate the cross-modal translation approach of learning a mapping from the SAR domain to the optical domain so that the translated output can serve as a visual complement when genuine optical imagery is unavailable. In the context of this study, the input SAR imagery was single-polarized (VV band, one channel), whereas the target optical imagery comprised three RGB channels, creating an asymmetry in channel dimensionality that must be handled by the generator architecture.

2.2. Convolutional Neural Network

A Convolutional Neural Network (CNN) is a deep learning architecture designed to process grid-structured data, such as images (LeCun et al., 1998). CNNs employ convolution operations to extract spatial features hierarchically: early layers learn low-level features, such

https://doi.org/10.35877/mathscience4878

ISSN : 2776-7922 (Print) / 2807-3037 (Online)

as edges, corners, and textures, whereas deeper layers learn high-level features, such as objects and complex structures. The three core components of a CNN are the convolutional, pooling, and fully connected layers.

The convolutional layer executes a convolution operation between the input image and learnable kernels (filters), producing feature maps that characterize the feature response at each spatial position. The pooling layer reduces the spatial dimension of the feature maps while providing invariance to small translations in the image (Goodfellow et al., 2016). The generator and discriminator architectures employed in this study incorporated convolution operations alongside normalization layers and nonlinear activation functions to construct effective image-processing blocks.

2.3. *Generative Adversarial Network*

The Generative Adversarial Network (GAN), as introduced by Goodfellow et al. (2014), comprises two neural networks trained simultaneously in a minimax game: a generator G and a discriminator D. The generator G transforms a random noise vector z, which is sampled from a prior distribution pz(z), into the data space to create synthetic samples G(z). The discriminator D differentiates between real samples from the data distribution and the synthetic samples generated by the generator.The GAN training framework is formulated as a two-player minimax game with the value function V(D, G), expressed as follows:

$$min_G max_D V(D,G) = \mathbb{E}_{S\sim p_{data}}[logD(s)] + \mathbb{E}_{z\sim p_z}[\log(1 - D(G(z)))] \quad (1)$$

where V(D,G) is the value function of the minimax game between the generator and discriminator, $E_{x\sim p\_data}[\log D(x)]$ is the expectation of the logarithm of the discriminator output over real samples x,D(x)∈[0,1] is the discriminator output representing the probability that x originates from real data, G(z) is the synthetic sample produced by the generator, and D(G(z)) is the discriminator output over the synthetic sample.

The discriminator is trained to maximize its ability to classify the real and synthetic samples. The discriminator loss is derived from the binary cross-entropy and is expressed as follows:

$$L_D = -\mathbb{E}_{S\sim p_{data}}[logD(s)] - \mathbb{E}_{z\sim p_z)}[\log(1 - D(G(z)))] \quad (2)$$

In the original minimax formulation, the generator is trained to minimize log(1 - D(G(z))). However, this approach frequently suffers from vanishing gradients when the discriminator is excessively strong in the early stages of training. Consequently, a non-saturating formulation was adopted.

$$L_G = -\mathbb{E}_{z\sim p_z)}[\log D(G(z))] \quad (3)$$

At the Nash equilibrium, the generator distribution p_g converges to the real data distribution p_{data}, and the optimal discriminator yields D*(x)=0,5 for all x, indicating that it can no longer distinguish between real and synthetic samples. In the image-to-image translation context, the generator input is replaced by an image from the source domain s, such that the output becomes G(s), a synthetic image in the target domain conditioned on the content of the input image.

2.4. *CycleGAN*

In many real-world cases, particularly in remote sensing, the availability of paired data is limited. CycleGAN (Zhu et al., 2020) enables image translation between domains without requiring paired datasets. The CycleGAN comprises two generators and two discriminators. The generator G:S→O translates an image from domain S (SAR) to domain O (optical), whereas generator F:O→S performs the inverse translation. Discriminator D_O distinguishes real optical images from translated images G(s), whereas discriminator D_S distinguishes real SAR images from back-translated images F(o).

https://doi.org/10.35877/mathscience4878

ISSN : 2776-7922 (Print) / 2807-3037 (Online)

The CycleGAN objective comprises two principal components: adversarial loss and cycle-consistency loss. The adversarial loss for mapping G: S → O is defined as follows:

$$L_{GAN}(G, D_O) = \mathbb{E}_{o\sim p_{data}(O)}[logD_O(o)] + \mathbb{E}_{s\sim p_{data}(S)}[\log(1 - D_O(G(s)))] \quad (4)$$

Analogously, the adversarial loss for mapping F:O→S is defined as:

$$L_{GAN}(F, D_S) = \mathbb{E}_{s\sim p_{data}(S)}[logD_S(s)] + \mathbb{E}_{o\sim p_{data}(O)}[\log(1 - D_S(F(o)))] \quad (5)$$

In practical implementations, CycleGAN employs the Least Squares GAN (LSGAN) formulation (Mao et al., 2017) to improve training stability by replacing the logarithmic function with a squared function, as shown in the following equation:

$$L_{LSGAN}(G, D_O) = \mathbb{E}_{o\sim p_{data}(O)}[(D_O(o) - 1)^2] + \mathbb{E}_{s\sim p_{data}(S)}[(D_O(G(s))^2] \quad (6)$$

The advantage of the LSGAN over the standard formulation lies in its non-vanishing gradients, even when synthetic samples lie far from the real data distribution, thereby mitigating the vanishing-gradient problem common to the cross-entropy formulation. A symmetric formulation is applied to the pair F and D_S.

The principal contribution of CycleGAN is the cycle-consistency loss, which ensures that a translated image can be consistently mapped back to its source domain, as expressed in the following equation:

$$L_{cyc}(G, F) = \mathbb{E}_{s\sim p_{data}(S)}\left[\| F(G(s)) - s \|_1\right] + \mathbb{E}_{o\sim p_{data}(O)}[\| G(F(o)) - o \|_1)] \quad (7)$$

The intuition behind cycle consistency is that if the translations G and F genuinely learn a meaningful cross-domain mapping, the composition of the two translations should approximate the identity function: F∘G ≈I_S and G∘F ≈ I_O. Without the cycle-consistency loss, infinitely many mappings could satisfy the adversarial loss alone; for instance, the generator could map all SAR images to a single optical image (mode collapse) as long as the image appeared realistic to the discriminator. The cycle-consistency loss reduces this solution space by requiring the mapping to be invertible (Zhu et al., 2020).

A third component that may be added is the identity loss defined in the following equation:

$$L_{identity}(G, F) = \mathbb{E}_{o\sim p_{data}(O)}\left[\| G(o) - o \|_1)\right] + \mathbb{E}_{s\sim p_{data}(S)}\left[\| F(s) - s \|_1)\right] \quad (8)$$

Identity loss was not applied in this study because the difference in channel dimensionality between the SAR domain (single-channel backscatter intensity) and optical domain (three-channel RGB reflectance) renders the input and target dimensionally incompatible, such that the expression G(o), where o belongs to the optical domain, is undefined for the mapping G:S→O. Accordingly, the overall CycleGAN objective adopted in this study is represented by the following equation:

$$L = L_{GAN}(G, D_O) + L_{GAN}(F, D_S) + \lambda L_{cyc}(G, F) \quad (9)$$

where λ_{cyc} is a hyperparameter that governs the relative contribution of the cycle-consistency loss to the total loss. In this study, λ_{cyc} = 10, following the configuration used by Zhu et al. (2020). The overall optimization objective is expressed as follows:

$$G^*, F^* = argmin_{\{G,F\}} max_{\{D_S, D_O\}} L\left(G, F, D_S, D_O\right) \quad (10)$$

The optimization is solved by alternating updates: at each iteration, the discriminator parameters are first updated to maximize L, after which the generator parameters are updated to minimize L iteratively until equilibrium is reached.

2.5. *GAN Training Stabilization*

A principal challenge in GAN training is the adversarial instability. When the discriminator becomes too strong relative to the generator, the gradient signal received by the generator

https://doi.org/10.35877/mathscience4878

ISSN : 2776-7922 (Print) / 2807-3037 (Online)

becomes uninformative, and learning is impeded. Conversely, if the discriminator is too weak, the generator receives insufficient competitive pressure to improve the quality of its output.

Heusel et al. (2017) introduced the Two Time-Scale Update Rule (TTUR), which assigns distinct learning rates to the generator and discriminator. A theoretical proof based on a two-time-scale stochastic approximation shows that the TTUR converges to a local Nash equilibrium under mild assumptions. The original TTUR generally sets a higher discriminator learning rate than that of the generator, so that the discriminator provides a more informative gradient signal. Nevertheless, the two-time-scale principle also permits the application of asymmetry in the opposite direction when the discriminator tends to dominate the training.

Another approach is spectral normalization (Miyato et al., 2018), which limits the Lipschitz constant of the discriminator by normalizing each weight matrix according to its spectral norm, as expressed in the following equation:

$$\bar{W} = \frac{w}{\sigma(W)} \tag{11}$$

where W is the weight matrix of a convolutional layer, σ(W) is the largest singular value of W obtained through singular value decomposition (SVD), and $\bar{W}$ is the normalized weight matrix. This approach is computationally lightweight, requires no additional hyperparameter tuning, and has been empirically shown to reduce the risk of mode collapse (Miyato et al., 2018). The application of spectral normalization to a CycleGAN architecture combined with data augmentation was reported by Fithriasari & Tjahjono (2026) for beach image translation, demonstrating improved training stability and generated image quality.

2.6. *Image Quality Evaluation*

Quality evaluation for unpaired image translation faces the challenge of the absence of optical reference images spatially synchronized with test SAR images. This study addresses that limitation by exploiting the cyclic nature of CycleGAN: full-reference metrics are computed between the input SAR image s and its cyclically reconstructed counterpart F(G(s)), that is, the result of the forward translation S→O followed by the backward translation O→S. This approach measures cycle reconstruction fidelity without requiring optical pairs, as applied in previous unpaired image translation studies (Fithriasari & Tjahjono, 2026). Four metrics were used: SSIM and LPIPS to measure cycle reconstruction fidelity; FID to measure the closeness of the generated-image distribution to that of the real optical imagery; and the convergence epoch to measure training speed and stability.

2.6.1. *SSIM*

The Structural Similarity Index (SSIM**)** evaluates the structural similarity of two images by analyzing luminance, contrast, and structural components (Wang et al., 2004), as expressed in the following equation:

$$SSIM(G(s), o) = \frac{(2\mu_{\{G(s)\}}\mu_o + C_1)(2\sigma_{\{G(s),o\}} + C_2)}{(\mu^2_{\{G(s)\}} + \mu^2_o + C_1)(\sigma^2_{\{G(s)\}} + \sigma^2_o + C_2)} \tag{12}$$

where μ_x and μ_y are the mean intensities, σ_x^2 and σ_y^2 are the variances, σ_{xy} is the covariance between the two images, and C_1 and C_2 are stabilizing constants. The SSIM ranges from [-1,1] to; values close to 1 indicate a high structural similarity.

2.6.2. *LPIPS*

The Learned Perceptual Image Patch Similarity (LPIPS) evaluates the perceptual difference between two images by analyzing their features within the space the feature space of a pre-trained neural network (R. Zhang et al., 2018), as expressed in equation:

$$\text{LPIPS}(x, y) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \| w_l \odot (\phi_l(x)_{hw} - \phi_l(y)_{hw}) \|_2^2 \tag{13}$$

https://doi.org/10.35877/mathscience4878

ISSN : 2776-7922 (Print) / 2807-3037 (Online)

where ϕ_l denotes the feature activation of layer l, w_l is the per-channel weight, and H_l and W_l are the spatial dimensions of that layer. A lower LPIPS value indicates a higher perceptual similarity.

2.6.3. *FID*

The Fréchet Inception Distance (FID) quantifies the difference between the feature distributions of generated optical images and real optical images within the Inception-v3 feature space, as represented by the following equation:

$$FID=\|\mu_r-\mu_g\|_2^2+Tr\left(\Sigma_r+\Sigma_g-2(\Sigma_r\cdot\Sigma_g)^{\frac{1}{2}}\right) \quad (14)$$

where μ_r, Σ_r and μ_g, Σ_g are the mean and covariance matrices of the features of the real optical images and the generated images, respectively, and Tr(·) is the trace operator. A lower FID value indicates a higher distributional similarity.

2.6.4. *Convergence Epoch*

The convergence epoch is defined as the first epoch at which the cycle-consistency loss falls below a threshold of 0.1 and remains stable below this threshold for the subsequent five epochs. This metric measures the speed and stability of the training convergence; an earlier convergence epoch indicates more efficient and stable training.

3. Research Method and Materials

3.1. *Data and Study Area*

The data analyzed comprised Sentinel-1 SAR imagery (VV band, GRD level, IW mode) and Sentinel-2 optical imagery (RGB bands, L2A level) for the island of Sumatra, acquired through the Google Earth Engine platform during the November 2025 flash-flood event. The training data consisted of 482 unpaired SAR images and 446 unpaired optical images taken from the Lhokseumawe, East Aceh, and Sibolga regions, whereas the test data consisted of 276 SAR images taken from the Aceh Tamiang region. All images were sampled as 256×256-pixel grids.

3.2. *Preprocessing*

SAR images in decibel (dB) units were converted to a linear scale via $p_{lin}=10^{(p_{db}/10)}$, then clipped to the range [0,1] and normalized to [-1,1] via $p'=2p_{lin}-1$. Optical images in reflectance units were directly clipped to [0,1] and normalized to [-1,1], respectively. This range adjustment is consistent with the Tanh activation function in the final layer of generator. An explicit despeckling stage (for example, a Lee filter or a learning-based filter) was not applied in this study; speckle reduction was expected to be learned implicitly by the generator during training, as reported in previous SAR-to-optical translation studies (Reyes et al., 2019).

3.3. *Model Architecture*

The generator adopts the ResNet-9blocks architecture (Zhu et al., 2020) comprising three stages: (i) a downsampling block with one 7×7 convolution followed by two 3×3 stride-2 convolutions, reducing the spatial resolution from 256×256 to 64×64 with 256 channels; (ii) nine residual blocks at the 64×64 bottleneck; and (iii) an upsampling block with two stride-2 transposed convolutions and a final 7×7 convolution terminating in the Tanh function. The generator G:S→O has one input channel and three output channels with a total of 11,371,907 trainable parameters. The generator F:O→S has a symmetric configuration (three input channels and one output channel).

The discriminator follows the PatchGAN design (Isola et al., 2017) with an effective 70×70 receptive field on the original input; at a resolution of 256×256, the discriminator output is a

https://doi.org/10.35877/mathscience4878

ISSN : 2776-7922 (Print) / 2807-3037 (Online)

30×30 matrix. All convolutional layers were followed by InstanceNorm2d and LeakyReLU activation (slope 0.2), except for the output layer. The total number of discriminator parameters is 2,764,737. In configuration B3, spectral normalization was applied to all the convolutional layers of the discriminator. The generator and discriminator architectures are presented in Tables 1 and 2, respectively.

Table 1: Generator Architectures

Layer Type	Output Size
Input Layer	(256,256,ch)*
ReflectionPad2D	(262,262,1)
Conv2D	(256,256,64)
InstanceNorm2D + ReLU	(256,256,64)
Conv2D	(128,128,128)
InstanceNorm2D + ReLU	(128,128,128)
Conv2D	(64,64,256)
InstanceNorm2D + ReLU	(64,64,256)
ResBlock 1-9	(64,64,256)
ConvTranspose2D	(128,128,128)
InstanceNorm2D + ReLU	(128,128,128)
ConvTranspose2D	(256,256,64)
InstanceNorm2D + ReLU	(256,256,64)
ReflectionPad2D	(262,262,64)
Conv2D	(256,256,3)*
Tanh	(256,256,3)

Note: *1 channel for G, 3 channels for F

Table 2: Discriminator Architectures

Discriminator without SN		Discriminator with SN	
Layer Type	Output Size	Layer Type	Output Size
Input Layer	(256,256,ch)*	Input Layer	(256,256,ch)*
Conv2D	(128,128,64)	SN-Conv2D	(128,128,64)
LeakyReLU	(128,128,64)	LeakyReLU	(128,128,64)
Conv2D	(64,64,128)	SN-Conv2D	(64,64,128)
InstanceNorm2D + LeakyReLU	(64,64,128)	InstanceNorm2D + LeakyReLU	(64,64,128)
Conv2D	(32,32,256)	SN-Conv2D	(32,32,256)
InstanceNorm2D + LeakyReLU	(32,32,256)	InstanceNorm2D + LeakyReLU	(32,32,256)
ZeroPadding2D	(34,34,512)	ZeroPadding2D	(34,34,512)
Conv2D	(31,31,512)	SN-Conv2D	(31,31,512)
InstanceNorm2D + LeakyReLU	(31,31,512)	InstanceNorm2D + LeakyReLU	(31,31,512)
ZeroPadding2D	(33,33,256)	ZeroPadding2D	(33,33,256)
Conv2D	(30,30,1)	SN-Conv2D	(30,30,1)

Note: *1 channel for D_S, 3 channels for D_O

3.4. *Experimental Configurations*

Seven training configurations were evaluated, designed to examine four aspects: training stabilization, isolation of the contributions of spectral normalization and augmentation, and sensitivity to the cycle consistency coefficient. where denotes instance normalization, denotes spectral normalization, and denotes the cycle consistency loss coefficient.

Table 3: Training configurations

Config	Gen. LR	Disc. LR	Disc. update	Disc. norm	Augmentation	λ
B1	2×10^{-4}	2×10^{-4}	every iteration	IN	—	10
B2	2×10^{-4}	5×10^{-5}	every 2 iterations	IN	—	10
B3	2×10^{-4}	1×10^{-4}	every 2 iterations	IN + SN	flip H/V	10
B4	2×10^{-4}	2×10^{-4}	every iteration	IN + SN	—	10
B5	2×10^{-4}	2×10^{-4}	every iteration	IN + SN	flip H/V	10
B6	2×10^{-4}	2×10^{-4}	every iteration	IN	—	5
B7	2×10^{-4}	2×10^{-4}	every iteration	IN	—	20

Configuration B1 replicates the default setting of Zhu et al. (2020). Configuration B2 adopts the two-time-scale principle by reducing the discriminator learning rate and its update frequency to restrain the discriminator dominance. Configuration B3 combines spectral normalization on all discriminator convolutional layers to control the Lipschitz constant, a

https://doi.org/10.35877/mathscience4878

ISSN : 2776-7922 (Print) / 2807-3037 (Online)

reduced discriminator learning rate of 1×10^{-4}, a reduced discriminator update frequency, and geometric augmentation.

Configurations B4 and B5 were designed to isolate the contribution of each component in B3. B4 applies spectral normalization alone, without augmentation, and without modifying the learning rate or update frequency (identical to B1 except for the addition of SN), whereas B5 adds geometric augmentation to B4 (the combination of SN and augmentation, but with a discriminator learning rate and update frequency identical to B1). Thus, the comparison of B1 → B4 → B5 isolates the effect of adding SN and augmentation without the confounding influence of learning rate changes. Configurations B6 and B7 explore the sensitivity of the cycle-consistency coefficient by setting λ=5 and λ=20, while keeping all other hyperparameters identical to B1.

Data augmentation was restricted to geometric transformations, namely, random horizontal and vertical flips. Photometric augmentation (variation of brightness, contrast, and saturation), commonly used in optical-to-optical translation (Fithriasari & Tjahjono, 2026), was not applied because the VV-polarized Sentinel-1 SAR imagery lacks color channels, so that saturation transformation is undefined, and because the variation in SAR intensity (backscatter in dB) has a different physical interpretation from the brightness variation in optical imagery.

3.5. *Training*

All convolutional weights were initialized using a normal distribution, $N(0, 0.02^2)$ ((Radford et al., 2016). Training employed the Adam optimizer (β_1=0.5, β_2=0.999) with a batch size of 1, a total of 200 epochs, and a linear learning rate decay schedule beginning at epoch 100. The cycle consistency loss coefficient was set to λ_{cyc}=10. An image buffer of size 50 was used to store the history of the generated images to stabilize the discriminator updates. Identity loss could not be applied to direction G:S→O owing to the difference in channel dimensionality between the source and target domains. The training was performed on an NVIDIA A100-SXM4-40GB GPU.

3.6. *Evaluation and Model Selection*

The quantitative evaluation employs the four metrics described in Subsection 2.6. Two full-reference metrics (SSIM and LPIPS) were computed between the original SAR image and the SAR image reconstructed through the cycle F(G(s)) over all 276 test images, thereby requiring no optical pairs. The FID was calculated by comparing the feature distributions of the generated images with those of real optical images. All metrics were evaluated at epoch-200 checkpoint.

The model selection was based on the convergence speed; the model that most rapidly attained training stability was deemed the best. The convergence epoch was defined as the first epoch at which the cycle-consistency loss fell below a threshold of 0.1 and remained stable for the subsequent five epochs. This criterion was chosen because, in the unpaired translation scenario, the speed and stability of convergence reflect the effectiveness of the training strategy in balancing the generator–discriminator dynamics. The relationship between the convergence epoch and the four quality metrics was analyzed to verify the consistency of these results.

4. Results and Discussion

4.1. *Quantitative Evaluation and Model Selection*

Table 4 presents the evaluation results for the seven configurations on the test set of 276 SAR images used in this study. Model selection was based on the convergence speed, as described in Subsection 3.6. SSIM and LPIPS were computed on the cyclic reconstruction F(G(s)); FID against the real optical distribution; and Conv. denotes the convergence epoch. The best values are highlighted with grey shading.

https://doi.org/10.35877/mathscience4878

ISSN : 2776-7922 (Print) / 2807-3037 (Online)

Table 4: Evaluation Results

Configurations	SSIM**	LPIPS*	FID*	Conv.*
B1	0.837 ± 0.006	0.084 ± 0.003	147.75	56
B2	0.904 ± 0.009	0.040 ± 0.003	164.54	50
B3	0.939 ± 0.006	0.025 ± 0.002	143.00	43
B4	0.875 ± 0.007	0.062 ± 0.002	164.08	53
B5	0.908 ± 0.007	0.048 ± 0.002	184.54	63
B6	0.787 ± 0.006	0.109 ± 0.002	153.68	72
B7	0.874 ± 0.007	0.056 ± 0.002	184.54	70

Note: Values after ± denote standard error (SE = Std/√n, n = 276), **the higher, the better *the lower, the better

The examination of the loss dynamics first revealed significant differences in the generator-discriminator balance. In the default configuration B1, the discriminator loss is very low at the end of training (0.057 and 0.093 for the SAR and optical domain discriminators, respectively), but rises by more than an order of magnitude on the test set (5.99 and 3.16, respectively), indicating that the discriminator overfits the training distribution and fails to generalize, thereby supplying a misleading gradient signal to the generator (Arjovsky et al., 2017). Reducing the discriminator learning rate and update frequency in B2 narrows the test/train gap, whereas the combination of spectral normalization and reduced discriminator dynamics in B3 yields both a balanced discriminator and the lowest cycle-consistency loss.

Based on the convergence speed criterion, configuration B3 was the best model, attaining training stability earliest at epoch 43, ahead of the other six configurations. This superiority is consistently reflected across all quality metrics: B3 achieves the highest SSIM (0.939), lowest LPIPS (0.025), and lowest FID (143.00). The consistency across both the convergence criterion and quality metrics reinforces the conclusion that B3 yields a model that converges the fastest while also achieving the highest quality. According to de Deijn et al. (2024), Pix2Pix achieved an FID of 174.43 on Cityscapes, whereas CycleGAN recorded an FID of 316.43 on Facades. Therefore, the FID of 143.00 achieved by B3 is quite competitive for cross-domain translation, especially considering the significant modality gap.

Configuration B2, designed to stabilize training by restraining the discriminator, yields good reconstruction fidelity (SSIM 0.904) but a worse FID than B1. The likely cause is that the B2 discriminator becomes too weak; reducing the discriminator learning rate to one-quarter and halving the update frequency simultaneously results in a less-sharp adversarial signal reaching the generator. This is consistent with the discussion by Heusel et al. (2017), who stated that the optimal two-time-scale ratio should not be too extreme. Spectral normalization in B3 offers a middle ground: the discriminator capacity is controlled through the Lipschitz constant constraint without prematurely weakening it via an aggressive reduction in the learning rate.

Furthermore, the cyclic reconstruction fidelity (SSIM, LPIPS) is not always aligned with the distributional quality (FID). For instance, B5 attains a high SSIM (0.908) and low LPIPS (0.048) but the worst FID (184.54), and B4 has a better SSIM than B1 (0.875 vs. 0.837) yet a worse FID (164.08 vs. 147.75). This is because SSIM and LPIPS, computed on the S→O→S cycle, can attain high values as long as the forward and backward mappings are mutually consistent, without guaranteeing that the intermediate optical image G(s) genuinely resembles the real optical distribution. According to Chu et al. (2017), CycleGAN can encode the source information within an intermediate image to ensure cycle consistency. Conversely, the FID assesses the congruence between the distribution of generated images and the distribution of real optical images, offering a complementary perspective. The optimal setup, B3, achieved excellence in both reconstruction accuracy (SSIM 0.939) and distributional similarity (FID 143.00), demonstrating its strength despite the limitations of each metric. This highlights the necessity of reporting both reconstruction and distribution-based metrics when assessing unpaired image translation.

To ensure that the observed superiority of B3 over the baseline B1 is not due to random seed selection, both configurations were retrained using two additional seeds. In three

https://doi.org/10.35877/mathscience4878

ISSN : 2776-7922 (Print) / 2807-3037 (Online)

independent trials, B3 consistently surpassed B1, achieving an SSIM of 0.933 ± 0.009 compared to 0.837 ± 0.014, an LPIPS of 0.031 ± 0.009 versus 0.082 ± 0.006, and an FID of 147.27 ± 4.95 versus 191.16 ± 37.67 (mean ± std across three runs). The minimal variance observed across runs suggests that the performance differences are systematic rather than attributable to a single advantageous initialization. The significantly reduced cross-run FID variance of B3 in comparison to B1 indicates that the application of spectral normalization, alongside diminished discriminator dynamics, results in more consistent distributional quality across various training trajectories.

4.2. *Ablation Study*

To determine the impact of each component in B3, two groups of ablation were conducted: one on spectral normalization and augmentation (B4, B5) and another on the cycle-consistency weight (B6, B7).

A comparative analysis of B1, B4 (spectral normalization only), and B5 (SN + augmentation) revealed that the integration of spectral normalization into a discriminator configuration characterized by aggressive updates at a full learning rate of 2×10^{-4} enhanced cyclic reconstruction fidelity. This is evidenced by an increase in the SSIM from 0.837 in B1 to 0.875 in B4 and 0.908 in B5. However, these configurations adversely affected the FID, with values deteriorating from 147.75 in B1 to 164.08 in B4 and 184.54 in B5, and resulted in slower convergence in B5, observed at epoch 63. These results indicate that the effectiveness of spectral normalization is not self-contained but depends on the accompanying discriminator learning rate and the update frequency. In B3, spectral normalization was combined with a reduced discriminator learning rate (1×10^{-4}) and a reduced update frequency (every two iterations), so that the three mechanisms synergistically maintained the generator–discriminator balance and yielded the best performance across all metrics. In contrast, applying spectral normalization without adjusting the discriminator dynamics (B4 and B5) is insufficient to relieve the discriminator dominance; thus, although cyclic reconstruction improves, the distributional quality of the generated images (FID) deteriorates.

The second group examines the cycle-consistency weight by comparing B6 ($\lambda = 5$) and B7 ($\lambda = 20$) with B1 ($\lambda = 10$). Reducing λ to 5 (B6) yields the worst performance on nearly all metrics (SSIM 0.787; LPIPS 0.109; slowest convergence at epoch 72) because relaxing the cycle-consistency constraint reduces the pressure to preserve cross-domain structural consistency. Increasing λ to 20 (B7) provides better reconstruction fidelity than B6 (SSIM 0.874) but a high FID (184.54) because the excessive weight on cyclic reconstruction suppresses the contribution of the adversarial loss, so that the generated images lose distributional realism. These results confirm that $\lambda = 10$, following the standard configuration of Zhu et al. (2020), is an adequate choice for this SAR-to-optical translation case and provides the basis for using $\lambda = 10$ in B3.

4.3. *Translations Results*

Visual examination of the translated images (Figure 1) shows that all seven configurations effectively mapped the SAR scenes to a tropical land-cover palette. In this palette, dark green signifies vegetation, dark tones represent water bodies, and brownish tones indicate flood-affected areas where soil and inundated land are exposed. The absence of saturated blue in the generated outputs aligns with the natural appearance of the Sentinel-2 imagery over the tropical study area during the November 2025 flood event. This color encoding is pertinent to flood monitoring because the spatial distribution of brown regions serves as a visual proxy for the extent of flood-affected land.

Configuration B3 provided the most coherent translations, reflecting its quantitative advantage. The distinctions between flood-affected regions, vegetation, and water bodies were clearly marked. The pattern of fields in the agricultural-aquaculture mosaic (row 1) is distinctly resolved, and the large winding river (row 2) is depicted as a sharply defined dark

https://doi.org/10.35877/mathscience4878

ISSN : 2776-7922 (Print) / 2807-3037 (Online)

channel against the surrounding flood-affected terrain. B1 and B2 generated similar results that also preserved the correct land–water polarity, although B2 exhibited slightly less overall contrast.

Conversely, configurations B4, B5, and B7 exhibited a notable inversion of the land and water features. In the first row, areas impacted by flooding, which appear distinctly bright in B1and B3, manifest as dark regions in B4, B5, and B7. Simultaneously, the channel network, typically represented as dark, is depicted as bright, land-like elements, effectively reversing the scene's interpretation and obscuring the spatial extent of the flood-affected area. This inversion is also apparent in row 2 for B7, where the land surrounding the winding river assumes light cream-yellow hues, thereby altering the contrast structure from its reference. This visual observation aligns with the quantitative data in Table 4, where B4, B5, and B7 maintain reasonable cyclic reconstruction fidelity (SSIM and LPIPS) but exhibit the poorest FID values among all configurations, indicating that the generated optical images significantly deviate from the actual Sentinel-2 distribution despite accurate cycle reconstruction. From the perspective of flood monitoring, the inversion phenomenon presents a distinct challenge, as it results in the misclassification of the category that the processed imagery is designed to emphasize: flood-affected areas.

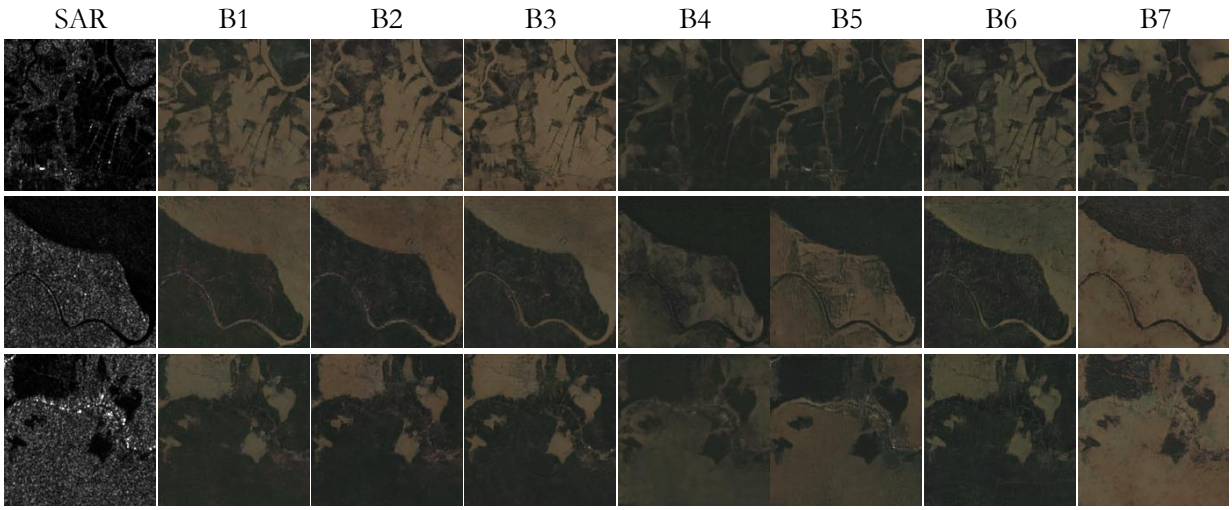


Figure 1: Translations Results

4.4. Limitation

This study has several limitations. The primary issue was the land–water inversion phenomenon observed in configurations B4, B5, and B7, where rivers appeared as bright land and flood-affected areas appeared as dark water. This effectively reversed the visual cues that the translated images were supposed to provide for flood monitoring. This inversion arises because the cyclic reconstruction metrics SSIM and LPIPS, computed over the S→O→S cycle, evaluate only whether the forward and backward mappings are mutually consistent and do not verify the semantic faithfulness of the intermediate optical image G(s). Therefore, a generator pair can achieve a high-fidelity reconstruction while encoding scene information in an inverted manner, consistent with the finding of Chu et al. (2017) that CycleGAN may conceal source information within the intermediate image to satisfy the cycle-consistency constraint. From the operational perspective of flood monitoring, this is particularly consequential because it misrepresents the flood-affected areas that the translated imagery is intended to highlight, compromising the practical utility of the model, even when its reconstruction metrics appear satisfactory.

SAR B4 B5 B7 SAR B4 B5 B7

https://doi.org/10.35877/mathscience4878

ISSN : 2776-7922 (Print) / 2807-3037 (Online)

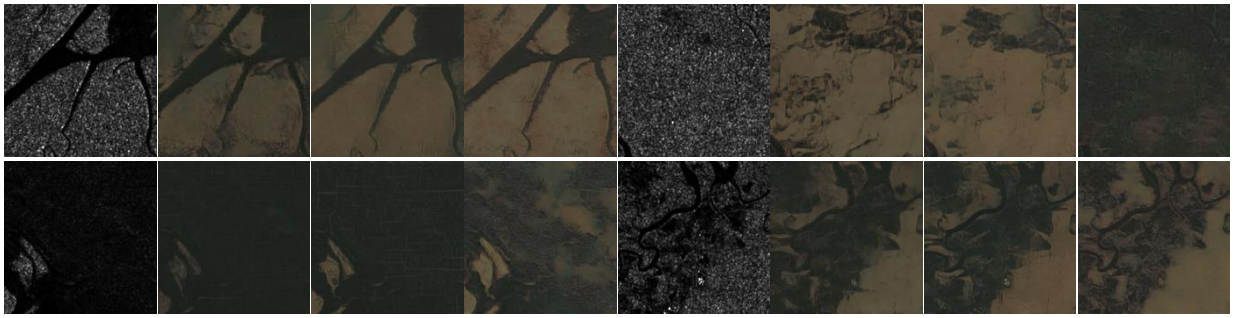


Figure 2: Semantic Inversions Cases

Several limitations should be acknowledged. The FID metric, which proved indispensable in revealing this inversion, was computed in an Inception-v3 feature space trained on ImageNet rather than remote sensing imagery; therefore, its sensitivity to geospatial content may not be optimal. The semantic correctness of the translation was assessed through visual inspection rather than against an independent land-cover reference; therefore, a formal quantitative measure of land–water class accuracy remains for future work. To evaluate the sensitivity of the results to random initialization, both the baseline (B1) and the optimal configuration (B3) were retrained using two additional random seeds, resulting in three independent runs for each configuration. The findings confirmed that the performance disparity between B3 and B1 remains consistent across these runs (refer to Section 4.1). The other five configurations (B2, B4-B7) were assessed based on a single run; expanding the multi-seed analysis to encompass all configurations is a potential avenue for future research. All configurations were trained at a resolution of 256×256; while higher resolutions might reveal further performance variations, they would necessitate increased computational resources.

5. Conclusion

This research explored the use of CycleGAN to SAR-to-optical translation on unpaired data from flood-affected areas of Sumatra, and compared seven training configurations evaluated using four metrics. Based on the convergence speed criterion, the configuration combining spectral normalization on the discriminator, a reduced discriminator learning rate, and geometric augmentation (B3) was the best model, attaining the earliest training stability (epoch 43) while also excelling across all quality metrics (SSIM 0.939; LPIPS 0.025; FID 143.00). Verification of multi-seed robustness has consistently demonstrated the superiority of B3 over the baseline across independent runs.

The isolated ablation study revealed that applying spectral normalization without adjusting the discriminator learning rate and update frequency improves cyclic reconstruction fidelity but worsens distributional closeness to the real optical imagery, indicating that the effectiveness of spectral normalization is not self-contained but depends on the accompanying discriminator dynamics. The inter-metric analysis further underscores the complementary role of the FID relative to the cyclic reconstruction metrics, as high reconstruction fidelity does not always guarantee distributional closeness to the real optical domain. The exploration of the cycle-consistency weight confirmed that $\lambda = 10$ was an adequate choice for this case.

Future research may explore various avenues. A semantic evaluation against an independent land-cover map could directly verify the accuracy of translations. Furthermore, employing higher-resolution outputs and partially supervised schemes offers promising approaches to address the land–water inversion observed in this study.

References

https://doi.org/10.35877/mathscience4878

ISSN : 2776-7922 (Print) / 2807-3037 (Online)

Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein generative adversarial networks. International Conference on Machine Learning, 214–223.

Chu, C., Zhmoginov, A., & Sandler, M. (2017). Cyclegan: A master of steganography. ArXiv Preprint ArXiv:1712.02950.

Fithriasari, K., & Tjahjono, B. K. (2026). Enhanced Beach Photo Translation using Modified Unsupervised GAN with Regularization. IIUM Engineering Journal, 27(1), 81–100. https://doi.org/10.31436/iiumej.v27i1.3824

Goodfellow, I. J., Bengio, Y., Courville, A., & Bengio, Y. (2016). Deep learning (Vol. 1, Number 2). MIT press Cambridge.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. Advances in Neural Information Processing Systems, 27.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium.

Isola, P., Zhu, J., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1125–1134.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278–2324.

Lee, J.-S., & Pottier, E. (2017). Polarimetric radar imaging: from basics to applications. CRC press.

Mao, X., Li, Q., Xie, H., Lau, R. Y. K., Wang, Z., & Paul Smolley, S. (2017). Least squares generative adversarial networks. Proceedings of the IEEE International Conference on Computer Vision, 2794–2802.

Miyato, T., Kataoka, T., Koyama, M., & Yoshida, Y. (2018). Spectral normalization for generative adversarial networks. ArXiv Preprint ArXiv:1802.05957.

Moreira, A., Prats-Iraola, P., Younis, M., Krieger, G., Hajnsek, I., & Papathanassiou, K. P. (2013). A tutorial on synthetic aperture radar. IEEE Geoscience and Remote Sensing Magazine, 1(1), 6–43.

Radford, A., Metz, L., & Chintala, S. (2016). Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. http://arxiv.org/abs/1511.06434

Reyes, M. F., Auer, S., Merkle, N., Henry, C., & Schmitt, M. (2019). SAR-to-Optical Image Translation Based on Conditional Generative Adversarial Networks — Optimization , Opportunities and Limits. 1–19. https://doi.org/10.3390/rs11172067

Torres, R., Snoeij, P., Geudtner, D., Bibby, D., Davidson, M., Attema, E., Potin, P., Rommen, B., Floury, N., & Brown, M. (2012). GMES Sentinel-1 mission. Remote Sensing of Environment, 120, 9–24.

Wang, L., Xu, X., Yu, Y., Yang, R., Gui, R., Xu, Z., & Pu, F. (2019). SAR-to-optical image translation using supervised cycle-consistent adversarial networks. Ieee Access, 7, 129136–129149.

Zhang, H., Li, H., Lin, J., Zhang, Y., Fan, J., Liu, H., & Liu, K. (2025). Seg-CycleGAN: SAR-to-optical image translation guided by a downstream task. IEEE Geoscience and Remote Sensing Letters, 22, 1–5.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 586–595.

Zhu, J., Park, T., & Efros, A. A. (2020). Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. 2223–2232.